Pearsonův test dobré shody chí kvadrát

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_{ei} - n_{oi})^2}{n_{oi}}$$

n_{ei} četnosti experimentální

n_{oi} četnosti očekávané (teoretické)

Test se nehodí pro soubory s velmi malými četnosti v jednotlivých kategoriích!!! Zde je vhodnější Kolmogorovův test.

Pearsonův i Smirnovův a Kolmogorovův test lze použít i pro diskrétní, ordinální i nominální data.

Pozor!!!

Všimněte si analogie mezi párovým testem (t-test, Wilcoxon) a testy shody (Smirnov, test dobré shody): v obou případech se sledují rozdíly mezi párovými hodnotami.

III.

Závislost

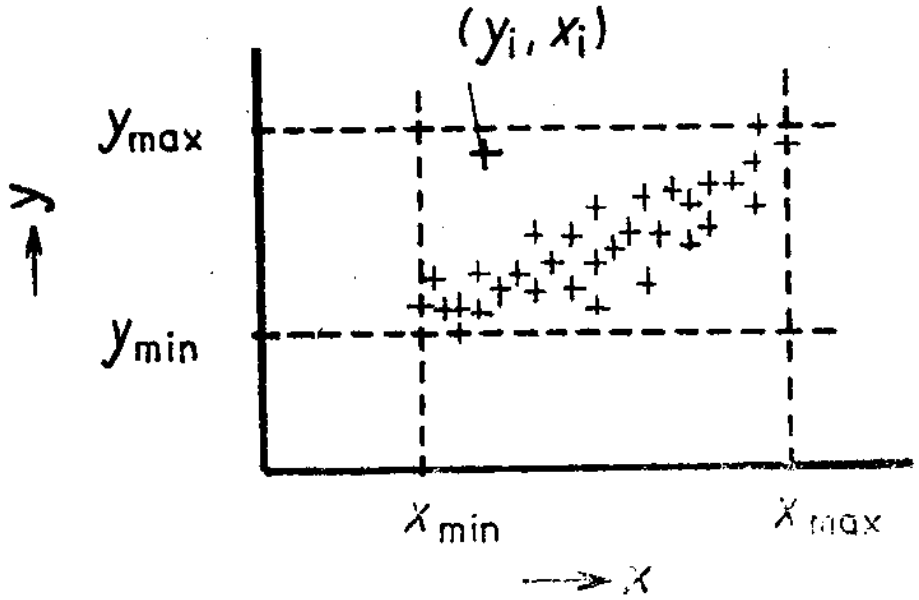
- Funkční
- Stochastická
  - Korelační
  - Regresní
    - » Lineární
    - » Nelineární

Závislost dvou proměnných

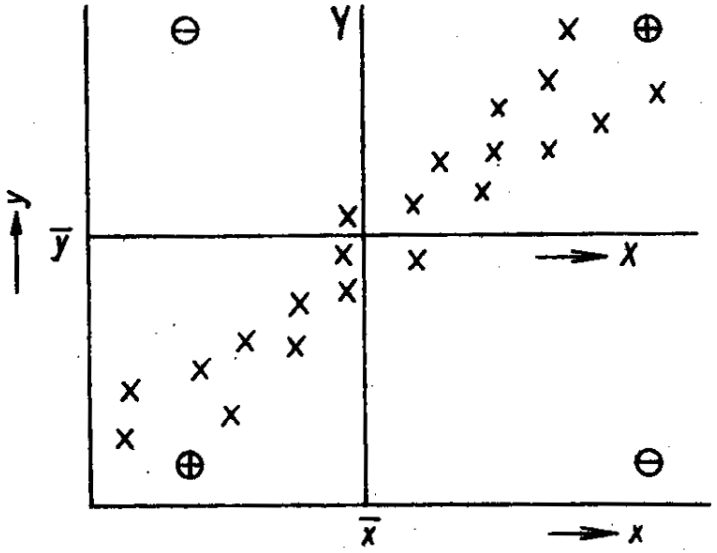
- Grafické nebo tabelární zobrazení dat
- Hledání základních konfigurací a tendencí v datech
- Výpočet numerických charakteristik

Závislost dvou proměnných

rozptylový graf (scatter plot)



kvadranty

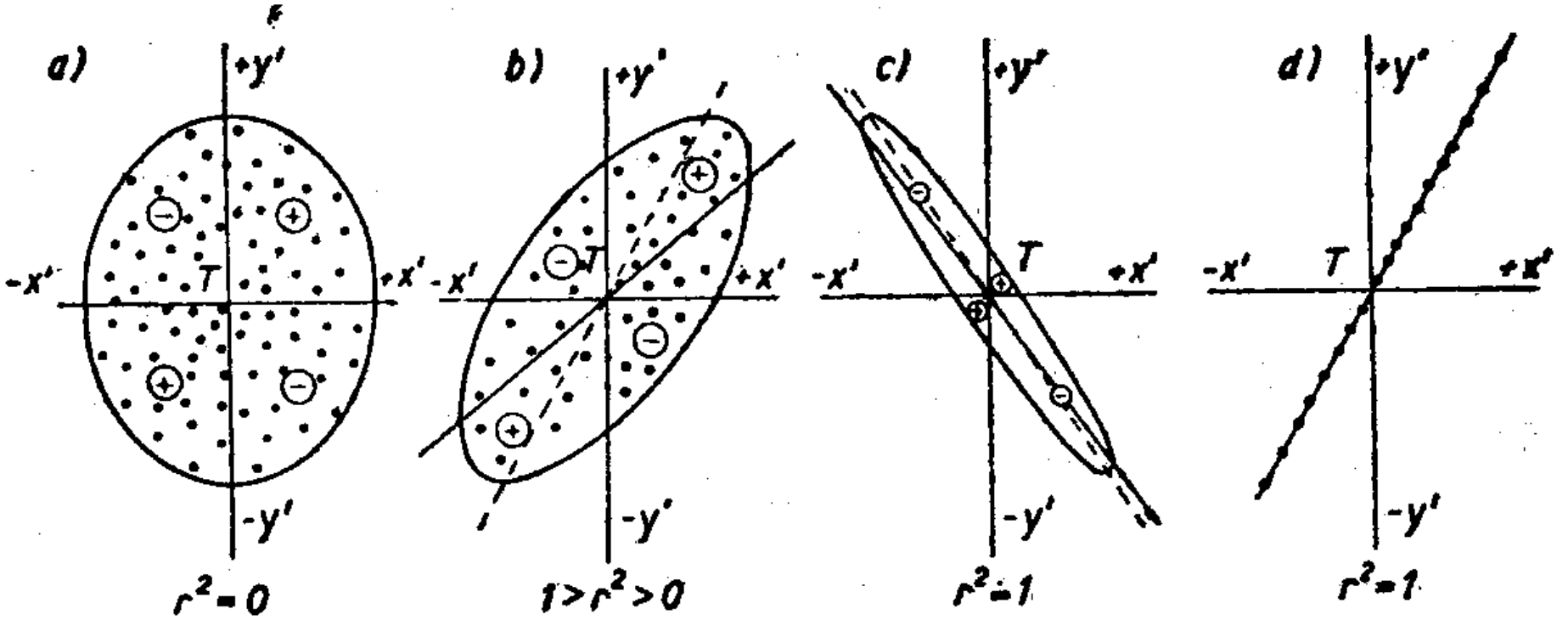


Geometrické znázornění korelace

X, Y – souřadnice těžiště,
$-$, $+$ negativní, resp. pozitivní kvadranty, v nichž jsou X_i a Y_i negativní, resp. pozitivní

Závislost dvou proměnných

Konfidenční elipsa pro danou hl. významnosti



Různé stupně korelace (znaménka kvadrantů platí pro součiny $x'y'$)

Závislost dvou proměnných

Konfidenční elipsa pro danou hl. významnosti

Střed elipsy: $0[\bar{x}, \bar{y}]$

Osa x svírá s delší osou elipsy úhel

$$tg2\alpha = \frac{2rs_x s_y}{s_x^2 - s_y^2}$$

Plocha elipsy:

$$Q = s_x s_y \sqrt{1 - r^2}$$

Pearsonův koeficient korelace

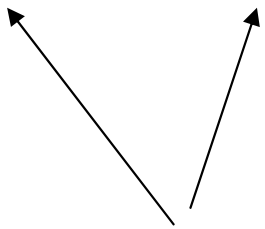
$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

kovariance

$0 \leq r \leq 1$

$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$



Standardizované hodnoty

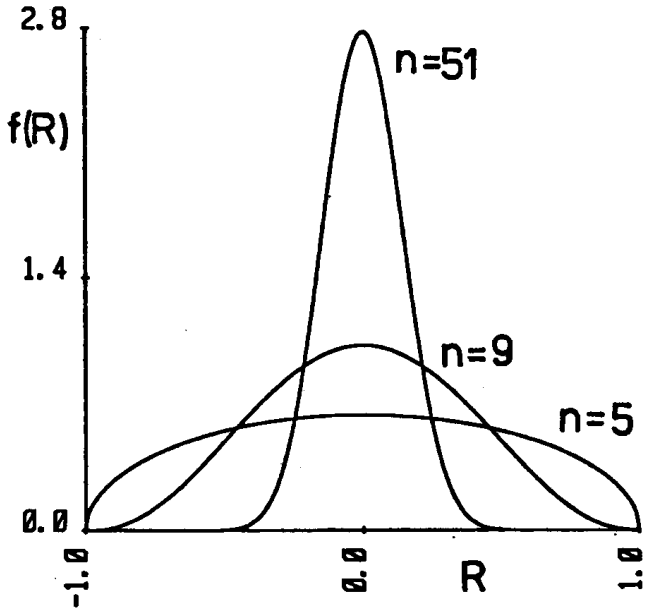
Pearsonův koeficient korelace

- Vyjadřuje pouze sílu lineárního vztahu.
- Je velmi ovlivněn odlehlými hodnotami.
- Nerozlišuje mezi závisle a nezávisle proměnnou.
- Obě proměnné musí mít náhodný charakter.
- **Korelace sama o sobě neznamená přítonmost příčinného vztahu!!!**

Odhad a testování Pearsonova k. k.

- H_0: r = 0

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$



Obr. 7.4 Hustota pravděpodobnosti výběrového korelačního koeficientu pro ρ = 0 a pro rozsahy výběru n = 5, 9, 51

-1 ≤ r ≤ 1

r^2 = koeficient determinace

Síla asociace /r/

Malá	0,1 – 0,3
Střední	0,3 – 0,7
Silná	0,7 – 1,0

Odhad a testování Pearsonova k. k.

- H_0: $r = \rho_0$

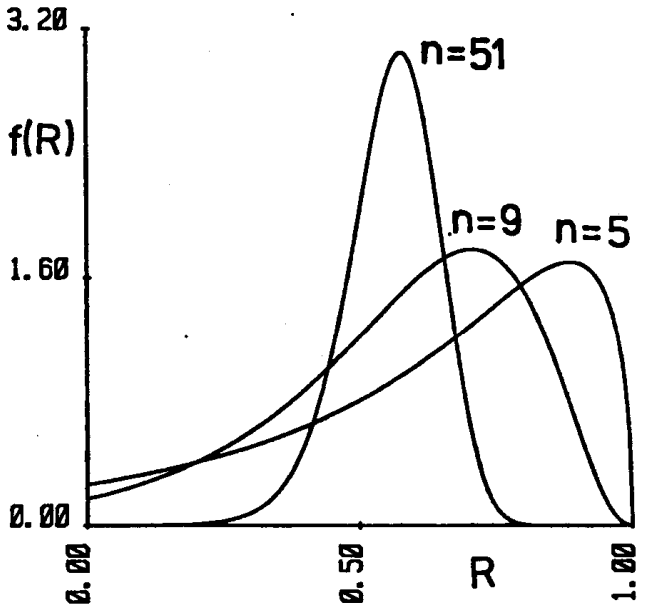
$$\mu_z = \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)$$

$$Z = \left|z_r - z_\rho\right|\sqrt{n-3}$$

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

$$s_z = \sqrt{\frac{1}{n-3}}$$

$$r = \frac{e^{2z}-1}{e^{2z}+1}$$



Obr. 7.5 Hustota pravděpodobnosti výběrového korelačního koeficientu (pro ρ = 0.6) pro rozsahy výběru n = 5, 9, 51

$$z - ts_z \le \mu_z \le z + ts_z$$

Odhad a testování Pearsonova k. k.

- H_0: $r_1 = r_2$

$$u = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

$$z_1 = \frac{1}{2} \ln\left(\frac{1 + r_1}{1 - r_1}\right)$$

$$z_2 = \frac{1}{2} \ln\left(\frac{1 + r_2}{1 - r_2}\right)$$

Pořadová korelace

- Spearmanův

$$r_s = 1 - \frac{6\sum D_i^2}{n(n^2-1)}$$

Hodí se spíše pro zařazovací ordinální data,
pro zařazovací ordinální data se však běžně používá.

D_i jsou rozdíly v pořadí hodnot x_i a y_i vzhledem k ostatním hodnotám výběru.

- Kendallův

Sleduje počet a charakter rozdílů v pořadí - pro j > i:

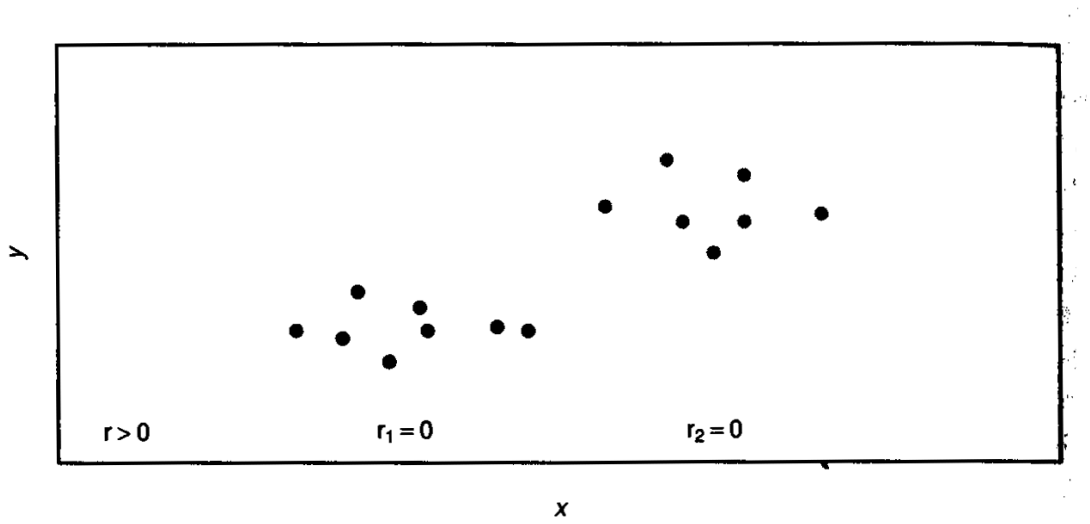
$y_j > y_i$ konkordance (kladná asociace) P
$y_j < y_i$ diskordance (kladná asociace) Q

$$t_k = \frac{P-Q}{n(n-1)/2}$$

Hodí se spíše pro porovnávací ordinální data

Druhy korelace

- Formální korelace: u percentuálních dat
- Korelace způsobená společnou příčinou
- Korelace způsobená nehomogenitou



Závislost a asociace nominálních dat

Kontingenční tabulky

čtyřpolní tabulka – pro dichotomická data

X	Y		Σ
	1	2	
1	n_{11}	n_{12}	$n_{1.}$
2	n_{21}	n_{22}	$n_{2.}$
Σ	$n_{.1}$	$n_{.2}$	n

Marginální četnosti

Asociace v kontingenčních tabulkách

Chí kvadrát test nezávislosti

$$\chi^2 = \frac{(ad-bc)^2 n}{(a+b)(c+d)(a+c)(b+d)}$$

$$\chi^2 = \frac{(|ad-bc| - n/2)^2 n}{(a+b)(c+d)(a+c)(b+d)}$$

m_i	o_j		$\sum$
	o_1	o_2	
m_1	a	b	$a+b$
m_2	c	d	$c+d$
$\sum$	$a+c$	$b+d$	$n = a+b+c+d$

Yatesova korekce na nespojitost

$$\Phi = \sqrt{\frac{\chi^2}{n}} = \frac{bc-ad}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

→ Je formálně shodný s Pearsonovým k. k.

Asociace v kontingenčních tabulkách

Fisherův exaktní (kombinatorický) test

– pro malé četnosti

$$P = \frac{(a+b)!(cd)!(a+c)!(b+d)!}{n!} \frac{1}{a!b!c!d!}$$

Zjišťujeme pravděpodobnost, že se vyskytne daná konfigurace četností a, b, c, d nebo jakákoli jiná, nulové hypotéze ještě nepříznivější(sloupcové a řádkové součty jsou stejné). Pokud je součet nižší než zvolená hladina významnosti – H_0 se zamítá.

Asociace v kontingenčních tabulkách

Woolfův G test nezávislosti

$$G = 2[a.\ln(a) + b.\ln(b) + c.\ln(c) + d.\ln(d) - (a + b).\ln(a + b) - (a + c).\ln(a + c) - (b + d).\ln(b + d) - (c + d).\ln(c + d)]$$

Pro výběry malého rozsahu se použije korekce na nespojitost dle Yatese:

je-li empirická četnost menší než teoretická přičteme 0,5

je-li empirická četnost větší než teoretická odečteme 0,5

Získaná hodnota se srovná s kritickou hodnotou rozdělení chí kvadrát pro (r-1)(s-1) stupně volnosti.

Koeficienty asociace

- Yule: $$Q = \frac{ac - bd}{ac + bd}$$
- Simple matching (pozorovaná shoda): $$SM = \frac{a + d}{a + b + c + d}$$
- Rusell – Rao: $$RR = \frac{a}{a + b + c + d}$$
- Rogers – Tanimoto: $$RT = \frac{a + d}{a + 2b + 2c + d}$$
- Sneath: $$DC = \frac{b + c}{a + b + c + d}$$

Koeficienty asociace

- Jaccard: $$J = \frac{a}{a+b+c}$$
- Kulczynski 1: $$K1 = \frac{a}{b+c}$$
- Sorensen – Dice: $$SD = \frac{2a}{2a+b+c}$$
- Anderberg: $$A = \frac{a}{a+2b+2c}$$
- Ochiai – Otsuka: $$O = \frac{a}{\sqrt{(a+b)(a+c)}}$$

Kontingenční tabulky

Y	Z				Σ
	1	2	...	c	
1	n_{11}	n_{12}	...	n_{1c}	$n_{1.}$
2	n_{21}	n_{22}	...	n_{2c}	$n_{2.}$
...	...	...	...	...	...
r	n_{r1}	n_{r2}	...	n_{rc}	$n_{r.}$
Σ	$n_{.1}$	$n_{.2}$	...	$n_{.c}$	n

Asociace nominálních dat

Chí kvadrát test nezávislosti

$$\hat{p}_{i.} = \frac{n_{i.}}{n} \qquad \hat{p}_{.j} = \frac{n_{j.}}{n}$$

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - n\hat{p}_{i.}\hat{p}_{.j})^2}{n\hat{p}_{i.}\hat{p}_{.j}} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}} = n\left[\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{n_{ij}^2}{n_{i.}n_{.j}} - 1\right]$$

Asociace v kontingenčních tabulkách

Koeficienty kontingence

- odvozené od koeficientu χ^2, pro čtyřpolní tabulky jsou shodné s koeficientem Φ.

Cramerův koeficient kontingence

$$K_1 = \sqrt{\frac{\chi^2}{n.\min(r-1, s-1)}}$$

Čuprovův koeficient kontingence

$$K_2 = \sqrt{\frac{\chi^2}{n.\sqrt{(r-1)(s-1)}}}$$

Asociace ordinálních kategorií

- Goodmanův – Kruskalův koeficient

$$\gamma = \frac{P - Q}{P + Q}$$

- Kendallovo tau-c

$$t_c = \frac{2m(P - Q)}{n^2(m - 1)}$$

kde m je menší z obou dimenzí v kontingenční tabulce.

Testy shody pro párová dichotomická data

Dichotomická data: Mc Nemarův test

$$\chi^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} = \frac{(b-c)^2}{b+c}$$

H_0: $p_{12} = p_{21}$

Pro malé četnosti (b + c < 30):

$$\chi^2 = \frac{(|n_{12} - n_{21}| - 1)^2}{n_{12} + n_{21}} = \frac{(|b-c| - 1)^2}{b+c}$$

Yatesova korekce na nespojitost

Před zásahem	Po zásahu		Σ
	+	−	
+	n_{11}	n_{12}	$n_{1.}$
−	n_{21}	n_{22}	$n_{2.}$
Σ	$n_{.1}$	$n_{.2}$	n

Před zásahem	Po zásahu		Σ
	+	−	
+	p_{11}	p_{12}	$p_{1.}$
−	p_{21}	p_{22}	$p_{2.}$
Σ	$p_{.1}$	$p_{.2}$	1

Dichotomická data: srovnání 2 metod

- Kappa koeficient shody

$$\kappa = \frac{Ao - Ae}{1 - Ae} = \frac{2(ad - bc)}{(a + c)(c + d) + (b + d)(a + b)}$$

Nominální data: Bowkerův test symetrie

Zobecnění McNemarova testu pro nominální znak s r úrovněmi:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{r} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}$$

Kritérium má přibližně chí kvadrát rozdělení s r(r-1)/2 stupni volnosti.